# 基于文本挖掘与复杂网络的我国绿色消费领域研究主题挖掘 *

刘杰平　徐金亚
（成都东软学院，成都　611844）

**摘　要:** [ **目的 / 意义** ]“发展绿色消费”是我国“十四五”规划和 2035 年远景目标纲要的重要内容之一。针对绿色消费领域研究主题的挖掘，有助于快速了解当前该领域的研究进展和热点，为进一步研究提供参考和指导。[ **方法 / 过程** ] 基于文本挖掘技术和复杂网络分析方法，提出“综合考虑文献标题、摘要和关键词，采用文本分词技术提取文献主题词，并基于 AHP 法确定二元主题词组共现权重”的方法；针对传统词频 g 指数无法有效排除“高频泛词”的情况，基于 TF-IDF 算法对传统词频 g 指数进行优化，提出 TI-g 指数；对 2010~2022 年我国绿色消费领域学术文献进行实证研究。[ **结果 / 结论** ] 绘制了 2010 年以来我国绿色消费领域研究主题演进热力图，并对 2018 年以来研究热点进行挖掘，识别出该领域研究的 4 大主题域。

**关键词:** 绿色消费　研究热点　文本挖掘　复杂网络　TF-IDF　词频 g 指数
**分类号:** TP391；G255
**DOI:** 10.31193/SSAP.J.ISSN.2096-6695.2023.03.08

## 0 引　言

绿色消费研究起源于 20 世纪 70 年代，1987 年，英国学者 Elkington 和 Hailes 将绿色消费定义为“为避免使用‘危害健康、资源浪费、过度包装、出自稀有动物或自然资源的商品，以及对他国，尤其是发展中国家不利的商品’的消费行为[1]”。世界环保组织则提出绿色消费的 5R 原则，即 Reduce, Reevaluate, Reuse, Recycle, Rescue[2]。严格来说，学术界并未对绿色消费的概念形成统一的定义。广义层面，一般认为绿色消费是在商品购买、使用和废置处理的全流程中产生的减少浪费、避免污染等行为；而狭义层面，绿色消费则更加侧重于绿色商品购买行为本身[3]。

* 本文系四川省社会科学重点研究基地四川省电子商务与现代物流研究中心课题“‘十四五’背景下电商消费者绿色购买意愿与行为研究”（项目编号：DSWL21-37）、中国高等教育学会年度规划重点课题“大数据专业知识图谱构建与智能问答平台研究”（项目编号： 22SZH0305）的研究成果之一。

[ 作者简介 ] 刘杰平（ORCID：0009-0002-7079-7089），男，系副主任，副教授，硕士，研究方向为大数据分析、数据挖掘，Email：liujieping@nsu.edu.cn；徐金亚（ORCID：0009-0007-2669-0483），男，副院长，副教授，博士，研究方向为数据库、动力系统、深度学习等，Email：xujinya@nsu.edu.cn。

随着我国公民绿色环保意识的不断加强，绿色消费理念得到广泛的传播和认可。根据《中国公众绿色消费现状调查研究报告》，83.34% 的受访者表示支持绿色消费行为，其中 46.75% 的受访者表示“非常支持”[4]。中国连续经营协会与阿里新服务研究中心研究认为，60% 以上的受访者知晓绿色消费，其中，00 后、90 后对绿色消费的认知明显高于其他年龄段，分别达 79% 和 70%[5]。在国务院 2013 年印发的《循环经济发展战略及近期行动计划》中，将“绿色消费”作为推进社会层面循环经济发展的一项重要措施。2014 年，李克强总理在国务院常务会议中提出要促进绿色消费，扩大节能产品生产[6]。至今，我国已出台了一系列与绿色消费相关的制度、计划、标准等。2021 年，“发展绿色消费”被写入“十四五”规划和 2035 年远景目标纲要中[7]，作为我国未来发展的战略之一。

随着公众绿色消费意识的增强以及各级政府对绿色环保的重视，学术界对绿色消费也进行了大量的研究，积累了大量的研究成果。为了更好地推动绿色消费研究，助力我国绿色消费战略实施，需要对该领域既往的研究成果进行综合分析，梳理近年来该领域的研究主题演进路线以及当前的研究热点，分析该领域研究存在的问题，以便更好地推动该领域的研究。本文研究发现，在绿色消费领域研究主题的分析方面，相关研究主要集中在理论探讨和定性分析上，缺少对研究主题现状和趋势的总结和定量分析[8]。此外，已有文献主题挖掘技术还存在研究对象单一、高频主题词选取方法主观或无法有效排除高频泛词等问题[9-13]。因此，本文基于文本挖掘技术和复杂网络分析方法，以 2010~2022 年我国绿色消费领域学术文献为研究对象，对该领域研究主题进行挖掘，梳理出该领域研究主题演进路线以及当前的研究热点，识别出该领域研究的 4 大主题域，并针对各主题域研究存在的不足，提出绿色消费领域未来可能的研究方向以及研究方法的改进。此外，针对同类文献主题挖掘技术存在的不足，提出了优化措施和改进建议，本文的研究思路、框架，以及关键技术，亦可为其他领域文献挖掘提供借鉴。

## 1　相关研究

复杂网络（Complex Network）源自 20 世纪 80 年代美国圣菲研究所（Santa Fe Institute，SFI）提出的复杂性科学领域[14]。复杂网络是由多个节点构成的高度复杂的关系网络，真实的复杂网络一般具有自组织和小世界等特性。复杂网络理论可以描述和研究复杂系统及其拓扑结构，自提出以来，复杂网络分析方法已被广泛应用于各种复杂系统研究[15]，如人才流动网[16]、交通网[17-18]、电力网[19-20]、金融网[21-22]、疾病传播[23]、舆情传播[24-26]、文献挖掘[27-29]等。

文献研究的主题域与复杂网络的社区特性类似，因此，复杂网络也被广泛应用于文献挖掘领域，成为文献挖掘的三大类方法之一[30, 31]。如 Holeab C 等提出一种基于语义和网络分析相结合的复杂文献挖掘方法，对面向未来的技术分析（Future-oriented Technology Analysis，FTA）学科的研究趋势进行了分析[32]。Wang Y 等基于复杂网络理论对 Scopus 数据库中的文献进行挖掘，定量描述了国际人才流动的显著特征[33]。Ortega J 等通过对 GSC（Google Scholar Citations）中的文献进行挖掘，发现美国在世界科学地图上占据主导地位[34]。Chae C 等对韩国人力资源

管理研究的语义网络结构进行关键词网络分析，表明韩国人力资源管理的整个网络结构具有复杂的社会建构语义结构[35]。辛娟娟等基于复杂网络的社区识别技术，对林业领域文献进行了研究，识别出八大主题研究领域[29]。刘俊楠等以测绘期刊为研究对象，对测绘领域研究热点进行了研究[10]。何波等基于复杂网络理论对中国经理人领域 28 年研究的演变趋势进行了研究[36]。在绿色消费方面，尽管有学者基于复杂网络分析方法对绿色消费理念传播等问题进行了研究，然而就绿色消费研究文献的挖掘非常少。社会网络分析是复杂网络相关知识在社会关系系统中的应用，刘永胜等基于社会网络的视角对我国绿色食品领域研究现状与趋势进行了分析[37]。杜先芸等基于社会网络分析和共词分析对我国绿色消费行为领域研究热点和主题趋势进行了研究[38]。

综上，复杂网络作为一种典型的文献挖掘方法，被广泛应用于各学科文献数据研究中。不过在绿色消费相关文献的挖掘中应用并不多。通过对相关研究文献的梳理，笔者认为，当前的研究还存在两方面不足：

一方面，绝大部分学者以文献关键词直接作为文献主题词进行文献挖掘[9-13]，而文献关键词具有主观性和语义模糊性[39]；也有学者将文献的标题、摘要、关键词等内容合并作为分析数据进行文献挖掘[28]，但是其将以上内容作为“整体”进行分析。笔者认为，标题、摘要、关键词等在反映文献主题时的重要性，即权重应有所差异。因此，本文提出应综合考虑文献标题、摘要和关键词，采用文本分词技术提取文献主题词，进一步基于 AHP 法（Analytic Hierarchy Process）确定二元主题词组的共现权重，该权重直接影响最终主题词网络构建时边的权重。

另一方面，为了排除大量低频主题词的干扰，学者们一般仅将高频主题词作为分析依据。而关于高频主题词数量的确定，不少学者依据经验确定[11, 13, 37, 40]，该方法缺乏理论指导，具有一定主观性。为了避免这种主观性的缺陷，杨爱青等基于学者影响力 g 指数提出了词频 g 指数，其核心思想是当且仅当研究主题的关键词总量 $N$ 中，有 $g$ 个关键词的累计频次不少于 $g^2$ 次，而 g+1 个关键词的累计频次少于 $(g+1)^2$ 次，此时的 g 为词频 g 指数[41]。可以看出，词频 g 指数的核心是以主题词出现的频次为依据。然而，根据文本挖掘领域的经典算法 TF-IDF 算法（Term Frequency-Inverse Document Frequency）的思想，在文本挖掘实践中，主题词中存在很多“高频泛词”，即出现频次虽然很高，但是其业务含义较弱。以本研究为例，仅从词频来看，“绿色消费”出现频次很高，但对于分析该领域的具体研究主题而言，该主题词并不能有效反映该领域的具体研究主题，即可视为高频泛词。鉴于此，本文基于 TF-IDF 算法，对词频 g 指数进行改进，并提出了 TI-g 指数，以弥补传统词频 g 指数可能存在高频泛词的不足。

## 2 研究思路与框架

本研究基于文本挖掘技术以及复杂网络、AHP 层次分析等理论，对绿色消费领域研究主题进行挖掘，研究思路与框架如图 1 所示。

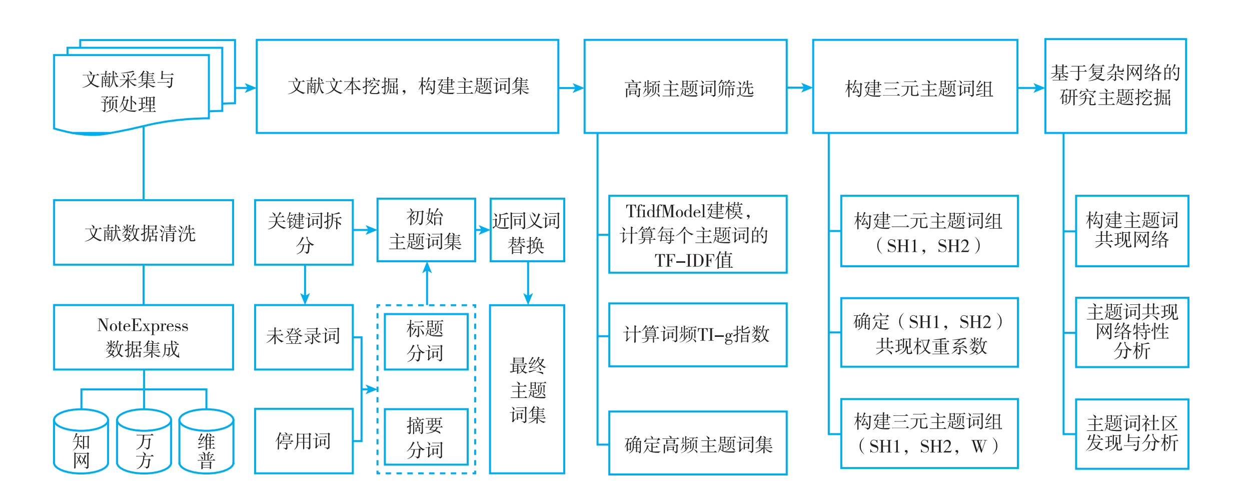



**图 1　研究思路与框架**

## 2.1　文献采集与预处理

本文在知网、万方、维普数据库中以“绿色消费”及“绿色购买”为关键词，采用“篇关摘”精确模式，检索该领域 2010 年 1 月 1 日至 2022 年 12 月 31 日发表的相关学术文献，并对文献数据进行合并、去重、格式规范化等处理，获得文献 7318 篇。随后，针对文献集中存在的部分特征缺失、重复及异常等问题，采用 Pandas 进行数据清洗，并剔除了新闻宣传、行业活动、征稿启事等非学术文献，再经过进一步人工核对，最终获得有效文献 4901 篇。

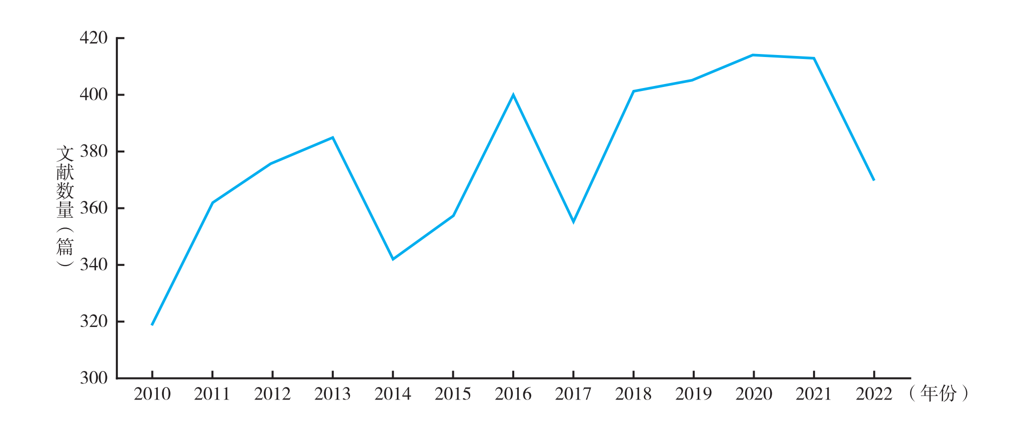



**图 2　2010~2022 年我国绿色消费领域文献数量**

年度发文数量如图 2 所示，可以看出，2010 年以来，我国绿色消费领域文献数量总体相对平稳，绿色消费领域一直是我国学者的研究重点之一。

## 2.2　文本分词及主题词提取

文献关键词本身即为相互独立的词语，而文献标题和摘要则需要进行文本分词。本文采用 jieba.lcut ( ) 方法对文献标题和摘要进行文本分词。在文本分词中，针对停用词，如“研究”“对

策”“建议”等，综合百度、哈工大、四川大学等停用词库及自定义停用词库进行过滤处理。针对未登录词，综合拆分后的关键词集以及主流输入法词库，如百度输入法、搜狗输入法、QQ 输入法等，以及自定义词等汇总形成未登录词集。最后，将近义词、同义词，如“大学生”和“高校学生”、“当代大学生”，“绿色消费行为”和“顾客绿色消费行为”等进行替换后，共获得 18007 个主题词。

## 2.3 高频主题词筛选

每个主题词在文献集中出现的频次不同，如“绿色消费理念”出现频次最高，为 4921 次。统计显示，前 3566 个主题词累计频次占比达到了 80%，而其余 14441 个主题词累计频次占比为 20%，且频次低于 6 次。因此，在实证研究中，并不需要对文献集中所有主题词进行研究，一般仅对高频主题词进行研究。

如前文所述，本文认为，高频主题词的选取可基于 TF-IDF 算法，对传统词频 $g$ 指数进行优化。TF-IDF 算法是一种文本挖掘加权技术，可有效避免仅以词频为基准来确定高频主题词，而出现无法排除高频泛词的问题。TF-IDF 算法的核心思想是，主题词在某篇文献中出现的频次越高，其权重越高；文献集中包含主题词的文献越多，其权重越低[42]。以文献集 $D=\{d_i \mid i=1,2,\ldots,n\}$ 为例， $W=\{w_j \mid j=1,2,\ldots,m\}$ 表示文献集 $D$ 的主题词集， $\bar{D}=\{w_j \in d_i \mid i=1,2,\ldots,n; j=1,2,\ldots,m\}$ 表示包含主题词 $w_j$ 的文献集，该主题词 $w_j$ 的 TF-IDF 值如式（1）所示。

$$(TF-IDF)_{w_j}=\frac{n_{w_j}}{n_{d_i}}*\lg\left(\frac{N_D}{N_{\bar{D}}+1}\right) \tag{1}$$

其中， $n_{w_j}$ 表示主题词 $w_j$ 在文献 $d_i$ 中出现的频次， $n_{d_i}$ 表示文献 $d_i$ 中主题词的总数， $N_D$ 表示文献集 $D$ 中文献的总数， $N_{\bar{D}}$ 表示文献集 $\bar{D}$ 中文献的数量。

结合 TF-IDF 算法及传统词频 g 指数思想，本文提出了一种基于 TF-IDF 算法的词频 g 指数计算方法的 TI-g 指数。TI-g 指数定义为：将所有主题词的 TF-IDF 值由高到低排序，当且仅当有 g 个主题词的累计 TF-IDF 值不少于 $g^2$ ，而 g+1 个主题词的累计 TF-IDF 值少于 $(g+1)^2$ 时，前 g 个主题词为文献集 $D$ 的高频主题词。TI-g 指数计算流程如下：

（1）采用式（1）计算主题词集 $W$ 中所有主题词 $w_j$ 的 TF-IDF 值。

（2）将主题词集 $W$ 中的主题词 $w_j$ 按照 TF-IDF 值降序排列，记 $w_j$ 的序号为 $r_j$ 。

（3）将主题词 $w_j$ 的 TF-IDF 值依次累加，主题词 $w_j$ 的 TF-IDF 累加和为 $\sum_{k=1}^{j}(TF-IDF)_{w_k}$ 。

（4）计算主题词 $w_j$ 的序号的平方，即 $r_j^2$ 。

（5）主题词 $w_j$ 的 TF-IDF 累加和与其序号平方相减，当二者差值的绝对值最小时，此时的序号 $r$ 为 TI-g 指数，即前 $r$ 个主题词为高频主题词。TI-g 指数公式如式（2）所示。

$$TI-g=\arg\min_x\left|\sum_{k=1}^{j}(TF-IDF)_{w_k}-r_j^2\right| \tag{2}$$

### 2.4　构建三元主题词组

基于复杂网络的文献研究主题挖掘本质上是一种共词分析，即将主题词（Subject Headings）两两组合，形成二元主题词组（$SH_1$，$SH_2$），然后再遍历（$SH_1$，$SH_2$）同时出现在文献 $d_i$ 中的频次 W，将该频次作为权重，形成三元主题词组（$SH_1$，$SH_2$，$W$），最终构建主题词网络。

如前文所述，对某个研究领域热点的挖掘是建立在单篇文献主题识别的基础上，一般以关键词这一单一要素为依据。但是，由于文献关键词由作者自行提出，具有主观性和语义模糊性，使得仅以文献关键词为基础的词频分析法和共词分析法存在一定的局限性[39]。鉴于此，本文将文献标题、关键词和摘要进行综合研究，避免仅以关键词为单一要素提取文献主题的局限性，使文献主题提取更为完整、可靠。

根据排列组合可知，主题词 $SH_1$ 和 $SH_2$ 在同一篇文献的标题、摘要和关键词中的共现情形有 6 种，不同的共现情形权重不同。为了确定主题词组在不同情形下同时出现的权重，本文采用 AHP 法，利用专门进行 AHP 分析的工具 yaahp 生成 AHP 调查软件，并邀请 12 位专家通过此软件，采用“1–9”标度法判断矩阵评分（专家评分一致性系数 CR 为 0.0176），确定了标题、关键词及摘要与文献主题之间的权重关系，分别为 0.47（$w_t$）、0.34（$w_k$）、0.19（$w_a$），并据此计算出二元主题词组（$SH_1$，$SH_2$）在不同情形下的共现权重系数，如表 1 所示。

**表 1　二元主题词组共现权重**

| 权重系数 | 权重系数的含义 | 权重系数计算 | 本文取值（归一化后） |
|---|---|---|---|
| $w_{t-t}$ | $SH_1$、$SH_2$ 同时出现在一篇文献的标题中 | $w_t*w_t$ | 0.38 |
| $w_{k-k}$ | $SH_1$、$SH_2$ 同时出现在一篇文献的关键词中 | $w_k*w_k$ | 0.26 |
| $w_{a-a}$ | $SH_1$、$SH_2$ 同时出现在一篇文献的摘要中 | $w_a*w_a$ | 0.13 |
| $w_{t-k}$ | $SH_1$、$SH_2$ 同时出现在一篇文献的标题和关键词中 | $w_t*w_k$ | 0.11 |
| $w_{t-a}$ | $SH_1$、$SH_2$ 同时出现在一篇文献的标题和摘要中 | $w_t*w_a$ | 0.07 |
| $w_{k-a}$ | $SH_1$、$SH_2$ 同时出现在一篇文献的关键词和摘要中 | $w_k*w_a$ | 0.05 |

因此，三元主题词组（$SH_1$，$SH_2$，$W$）的权重 $W$ 如下式（3）所示，其中 $n$ 为共现频次。

$$W=\sum w_i*n_i\left(i=t,t;k,k;a,a;t,k;t,a;k,a\right) \tag{3}$$

### 2.5　基于复杂网络的研究主题挖掘

文献热点不是由单个主题词构成，而是由一组紧密连接的点组成，这与复杂网络中社区的概念相似[29]。因此，可以使用复杂网络中社区发现算法挖掘文献热点。通过对复杂网络小世界特性及无标度特性的分析，可以了解复杂网络的特征，模块度 Q 值可以评价社区划分的优劣。

网络的平均聚类系数越大且平均路径长度越小，则网络的小世界特性越明显。在实际分析中，一般与相同规模的随机网络进行比较，进而判断网络是否具备小世界特性。$C_a$、$L_a$ 分别表示实际网络的平均聚类系数、平均路径长度；$C_r$、$L_r$ 分别表示相同规模随机网络的平均聚类系数、

平均路径长度。如果满足式（4）大于1，则认为实际网络具备小世界特性，且式（4）的值越大，则小世界特性越明显[43]。

$$\left(C_a \big/ L_a\right) \Big/ \left(C_r \big/ L_r\right) \tag{4}$$

用 $P(k)$ 表示网络中度为 $k$ 的节点出现的频率，如果 $P(k)$ 服从幂律分布，则网络具有无标度特性。该特性强调网络节点间资源的不平等分配，式（5）中幂指数 $\gamma$ 通常取值为 2~3。

$$P(k) \propto k^{-\gamma} \tag{5}$$

模块度 Q 常用于评价社区划分的优劣，Q 的值越接近于 1，表明网络的社区结构越好，具有明显社区结构的 Q 值一般在 0.3~0.7 之间[44]。

$$\mathrm{Q} = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(i,j) \tag{6}$$

式（6）中，$m$ 表示网络中边的总数，即节点间边权重总和；$A_{ij}$ 是由节点 i 和 j 之间边权重构成的邻接矩阵；$k_i$ 和 $k_j$ 分别表示节点 $i$ 和节点 $j$ 的度；当 $i$ 和 $j$ 处于同一个社区时，$\delta(i,j)$ 为 1，否则为 0。

## 3　我国绿色消费领域研究主题挖掘

### 3.1　我国绿色消费领域研究主题演进

基于本文提出的 TI-g 指数，对 2010~2022 年期间绿色消费领域的学术文献进行高频主题词挖掘，并通过热力图对研究主题的演进情况进行可视化展示，如图 3 所示。

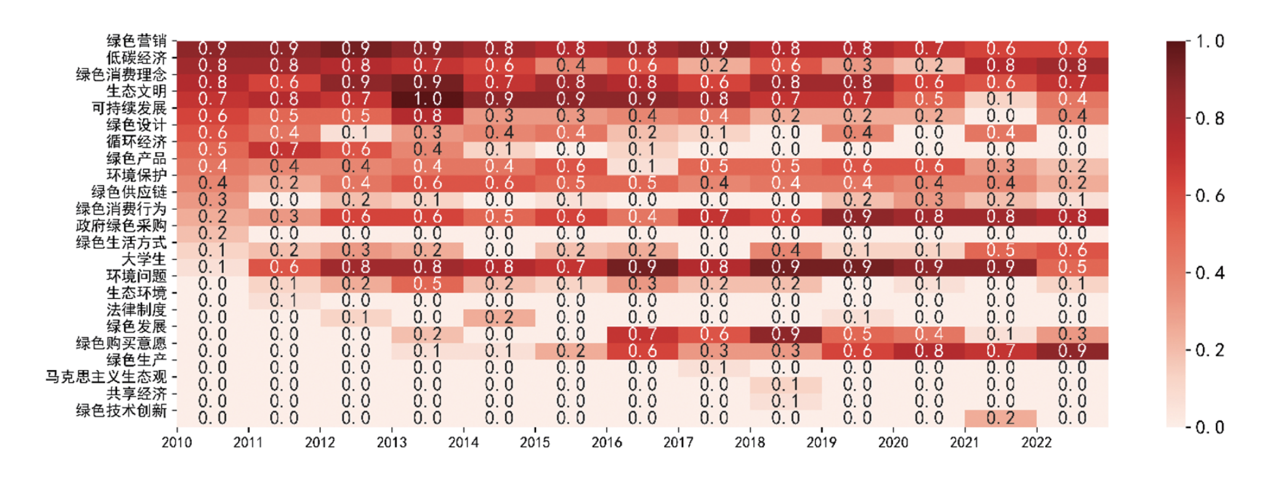



**图 3　2010~2022 年我国绿色消费领域研究主题演进**

总体而言，绿色营销、低碳经济、绿色消费理念、生态文明等一直是我国绿色消费领域研究的热点。此外，2016 年以前，我国绿色消费领域研究主要以基础理念为主（图 3 左上），研究立意多与环境相关，如可持续发展、循环经济、低碳经济、环境保护、环境问题以及法律制度等。2016 年以后，相关研究明显偏向于实务（图 3 右下），如绿色发展、特定群体（如大学生）绿色购买意愿、绿色购买行为的影响因素等心理研究，以及绿色设计、绿色生产、绿色技术创新等技术研究。

## 3.2　我国绿色消费领域研究热点主题

为深入挖掘近年来我国绿色消费领域的研究热点，在高频主题词挖掘基础上，基于复杂网络分析方法，构建了近五年（2018~2022）我国绿色消费领域研究主题词网络。主题词网络为无向图网络，详细网络指标如表 2 所示。

**表 2　主题词网络指标**

| 指标 | 节点 | 非零边 | 平均度 | 平均加权度 | 图密度 | 模块化 | 平均聚类系数 | 平均路径长度 |
|---|---|---|---|---|---|---|---|---|
| 数值 | 634 | 20751 | 65.461 | 30.091 | 0.103 | 0.289 | 0.614 | 1.906 |

### 3.2.1　主题词网络特性分析

（1）小世界特性分析

根据表 2，主题词网络平均聚类系数 $C_a$ 和平均路径长度 $L_a$ 分别为 0.614、1.906，而同等规模下随机网络的平均聚类系数 $C_r$ 和平均路径长度 $L_r$ 的等阶量分别为 0.104、1.898。根据式（4）计算可知，5.879>>1。因此，该主题词网络具备小世界特性，即任意两个节点之间的距离都比较短，同时主题词网络中存在着一些紧密相连的社区或主题域。

（2）无标度特性分析

主题词网络节点度的分布如图 4 所示，在双对数坐标系中，节点度分布的线性关系很弱，线性判定系数 $R^2$ 仅为 0.292，说明节点度的分布不服从幂律分布，即不存在大量节点具有较小度值，而少量节点具有很大的度值。根据式（5）可知，主题词网络不具备无标度特性，即主题词网络中大多数节点连接数相当，没有明显的“超级节点”。

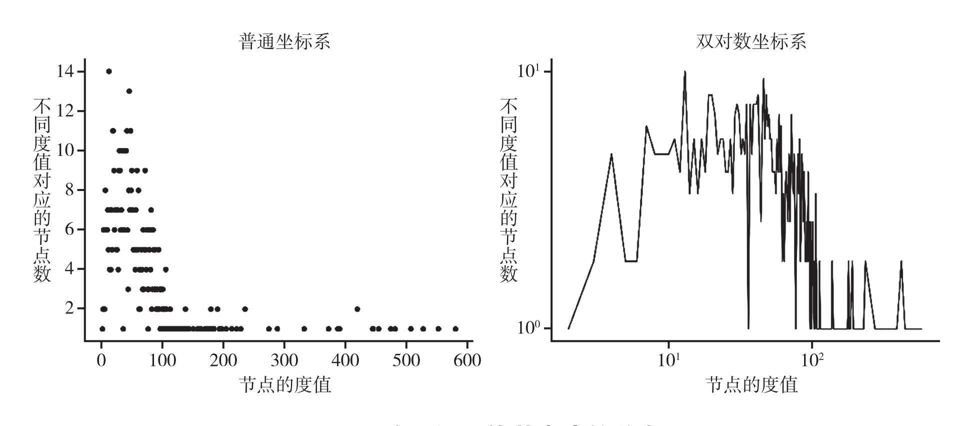



**图 4　主题词网络节点度的分布**

### 3.2.2 主题词网络社区发现与分析

本文采用 Gephi 进行社区划分，并采用 OpenOrd 算法进行可视化布局。在标准解析度下，共划分出 4 个社区（用 4 种颜色表示），如图 5 所示。基于 Gephi 默认的 Blondel V D 等人提出的算法[45]，社区划分优劣评价指标模块度 Q 值为 0.289，根据社区可视化效果以及式（6）可知，主题词网络不具备明显的社区结构，各研究主题域的交叉融合较多，没有形成相对清晰、独立的研究体系和方向。尽管如此，总体上仍可以将绿色消费领域研究热点分为 4 个主题域，各主题域核心主题词如表 3 所示。

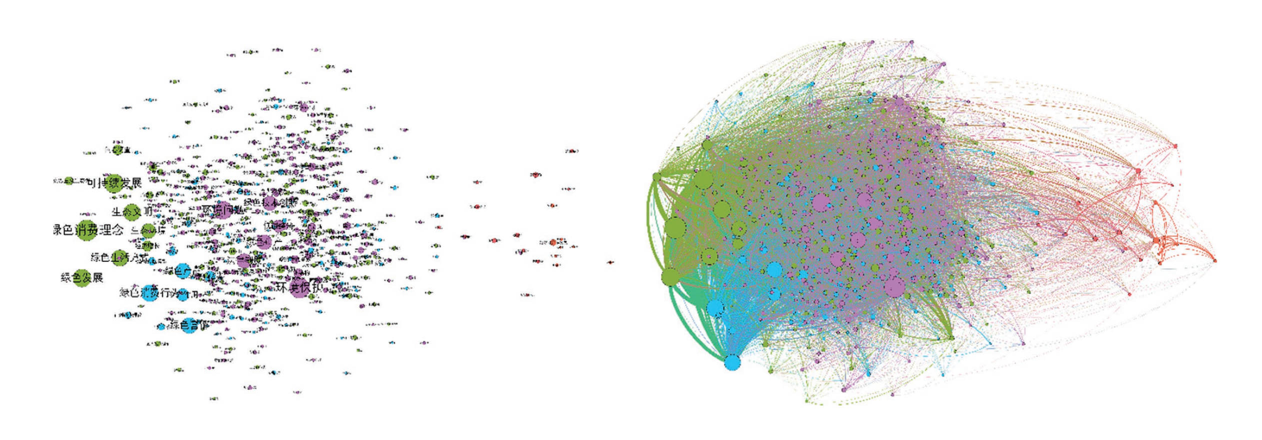



**图 5 主题词网络社区可视化全景**

**表 3 各研究主题域核心主题词**

| 序号 | 主题域占比 | 核心主题词（节点度大小排名前 15-20） | 主题域定义 |
| --- | --- | --- | --- |
| 1 | 22.29% | 绿色消费行为、绿色营销、绿色产品、中介作用、影响因素、结构方程模型、回归分析、计划行为理论、感知价值、收入水平、主观规范、消费者环保意识、因子分析、绿色消费态度、感知行为控制、宣传力度 | 绿色消费驱动因素研究 |
| 2 | 32.01% | 绿色消费理念、可持续发展、绿色发展、生态文明、绿色生活方式、生态环境、经济增长、马克思主义生态观、大学生、绿色消费教育、建设美丽中国、绿色文化、五大发展理念、乡村振兴、宣传教育 | 绿色消费价值观培育研究 |
| 3 | 42.52% | 环境保护、环境问题、绿色生产、法律制度、绿色技术创新、低碳经济、产业结构、政府绿色采购、绿色供应链、绿色创新、绿色产业、绿色设计、创新能力、政策支持、公众参与、评价指标、博弈模型、激励机制、环境绩效 | 绿色消费相关制度机制研究 |
| 4 | 3.18% | 经济可持续发展、服务消费、双循环、消费结构、绿色农业、绿色消费信贷、旅游消费金融、知识产权 | 其他研究 |

根据各主题域核心主题词，将其依次定义为：绿色消费驱动因素研究、绿色消费价值观培育研究、绿色消费相关制度机制研究以及其他研究。

绿色消费驱动因素研究主要是研究消费者的绿色消费行为、绿色生活方式是如何形成的，其驱动因素有哪些。驱动因素既包括内因，如感知价值、收入水平、消费者环保意识等，也包括外

因，如主观规范、宣传力度等。

绿色消费价值观培育研究主要是研究消费者，尤其是特定消费群体，如大学生、高中生等未来消费主体，对绿色消费、人与自然等理念的认知及宣传、教育、培养等，其间融合了马克思主义生态观、思想政治教育、乡村振兴等概念。进一步研究文献发现，该领域研究更多侧重于价值观培育的意义、形式、方法等层面，而对于培育的效果评价研究较少。

绿色消费相关制度机制研究主要是研究绿色消费与环境的关系，绿色消费相关的法律制度、税收制度、评价指标、激励机制以及相关技术创新机制等。

其他研究则是绿色消费理念在其他领域的应用研究，如绿色消费信贷、绿色消费金融等。

## 4 总 结

本文基于文本挖掘技术和复杂网络分析方法，进行了针对性的优化创新后，对 2010 年以来我国绿色消费领域研究主题演进以及 2018 年以来绿色消费领域研究热点的挖掘，可为其他学者进一步研究提供参考。

在研究方法层面，针对同类研究大多直接以文献关键词作为文献主题词存在的主观性和语义模糊性，本文提出“综合考虑文献标题、摘要和关键词，采用文本分词技术提取文献主题词，并基于 AHP 法确定二元主题词组共现权重”的研究方法。针对传统词频 g 指数无法有效排除“高频泛词”的情况，本文基于 TF-IDF 算法对传统词频 g 指数进行优化，提出 TI-g 指数，弥补了传统词频 g 指数的不足。

在实证研究层面，针对绿色消费领域研究主题，本文认为：

（1）2016 年以前，我国绿色消费领域研究主要以基础理念为主，研究立意更多与环境相关，如可持续发展、循环经济、环境保护、环境问题以及法律制度等。2016 年以后，相关研究则更明显偏向于实务，如绿色经济发展、特定群体（如大学生）绿色购买意愿、绿色购买行为相关的影响因素、绿色设计、绿色生产与绿色技术创新。

（2）2018 年以来，我国绿色消费领域研究总体可分为绿色消费驱动因素研究、绿色消费价值观培育研究、绿色消费相关制度机制研究以及其他研究四个主题域。但是，根据复杂网络分析方法，以上四个主题域社区结构不明显，且不具有明显的无标度性。因此，宏观来看，未来各主题域应该加强研究深度，在研究方法和跨学科研究上做更多探索，形成更加完善的研究体系。

（3）微观来看，各主题域研究还存在一定研究不足，例如：在绿色消费驱动因素研究主题域中，研究方法、理论和模型过于统一，如大部分研究均使用 SEM 结构方程模型或其他传统统计学方法，缺少如文本挖掘、情感分析、舆情分析等更为前沿的大数据相关技术进行研究；绿色消费价值观培育研究中，基于人群特征的差异化绿色消费价值观培育方式和评价方式研究不足；此外，如何形成和完善绿色消费相关的法律法规、税收政策等体制机制问题，以及绿色技术、绿色生产、绿色设计等实务问题，同样需要深入研究。

## 【参考文献】

［1］Elkington J, Hailes J. The green consumer guide: from shampoo to champagne-highstreet shopping for a better environment［M］. London: V. Gollancz, 1988.

［2］崔巧环. 我国施行绿色消费的影响因素及对策分析［J］. 理论导刊，2007（10）：116-118.

［3］侯海青，尹丽. 绿色消费行为影响因素研究综述［J］. 中国经贸导刊（中），2021（6）：149-152.

［4］中国生态文明研究与促进会. 中国公众绿色消费现状调查研究报告［R］. 2019.

［5］中国连锁经营协会，阿里新服务研究中心. 迈向新服务时代——生活服务业数字化发展报告（2021）［R］. 2021.

［6］马维晨，邓徐. 我国绿色消费的政策措施研究［J］. 环境保护，2017，45（6）：56-59.

［7］中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要［M］. 北京：人民出版社，2021.

［8］刘永胜，甘莹莹. 我国绿色食品领域研究现状与趋势分析——基于社会网络的视角［J］. 中国农业资源与区划，2020，41（2）：26-34.

［9］高阳，熊巨华，吴浩，等. 2021年度自然科学基金申请书关键词透视地理科学研究前沿热点与发展方向［J］. 地理科学，2022，42（1）：15-30.

［10］刘俊楠，刘海砚，陈晓慧，等. 测绘领域研究热点可视化分析［J］. 科学技术与工程，2019，19（32）：43-51.

［11］郑伟，刘玉林. 基于复杂网络的高校教师职业倦怠热点研究［J］. 黑龙江高教研究，2020，38（6）：50-55.

［12］谭章禄，彭胜男，王兆刚. 基于聚类分析的国内文本挖掘热点与趋势研究［J］. 情报学报，2019，38（6）：578-585.

［13］张若凡，申怡然，刘泽华. 基于复杂网络的中国管理学领域研究热点及演进［J］. 统计与管理，2019（8）：104-109.

［14］戴汝为，李耀东. 基于综合集成的研讨厅体系与系统复杂性［J］. 复杂系统与复杂性科学，2004（4）：1-24.

［15］姜敏勤，石小晶，姚安阳，等. 基于CSSCI数据的复杂网络研究热点、知识演进与趋势探析［J］. 武汉商学院学报，2022，36（2）：74-79.

［16］靳军宝，曲建升，吴新年，等. 中国高层次科技人才省际流动复杂网络特征研究［J］. 科技管理研究，2021，41（21）：112-118.

［17］Han P, Yinzhen L, Changxi M. Topology analysis of Lanzhou public transport network based on double-layer complex network theory［J］. Physica A: Statistical Mechanics and its Applications,2022(592): 126694.

［18］胡军，王雨桐，何欣蔚，等. 基于复杂网络的全球航空网络结构分析与应用［J］. 计算机科学，2021，48（S1）：321-325.

［19］Overbye T J, Shetye K S, Hutchins T R. Power grid sensitivity analysis of geomagnetically induced currents［J］. IEEE Transactions on Power Systems, 2013(28):2821-2828.

［20］傅杰，邹艳丽，谢蓉. 基于复杂网络理论的电力网络关键线路识别［J］. 复杂系统与复杂性科学，2017，14（3）：91-96.

［21］吴德胜，曹渊，汤灿，等. 分类管控下的债务风险与风险传染网络研究［J］. 管理世界， 2021，37（4）：35-54.

［22］谢赤，贺慧敏，王纲金，等. 基于复杂网络的泛金融市场极端风险溢出效应及其演变研究［J］. 系统工程理论与实践，2021，41（8）：1926-1941.

[ 23 ] Angela L, Nicola A, Alfonso M, et al. Complex network modelling of origin–destination commuting flows for the COVID–19 epidemic spread analysis in Italian lombardy region [ J ]. Applied Sciences,2021,11(10):4381.

[ 24 ] Mukul G, Rajhans M. Spreading the information in complex networks: Identifying a set of top–N influential nodes using network structure [ J ]. Decision Support Systems,2021(149): 113608.

[ 25 ] 贾红雨，赵雪燕，邱晨子. 基于复杂网络的微博网络舆情图谱分析方法研究 [ J ]. 现代情报，2015，35（3）：64–67.

[ 26 ] 彭程，祁凯，黎冰雪. 基于SIR–EGM模型的复杂网络舆情传播与预警机制研究 [ J ]. 情报科学，2020，38（3）：145–153.

[ 27 ] 孙海生. 基于Web of Science数据库的文献耦合网络实证研究 [ J ]. 情报杂志，2018，37（10）：201–207.

[ 28 ] 王燕鹏，韩涛，陈芳. 融合文献知识聚类和复杂网络的关键技术识别方法研究 [ J ]. 图书情报工作，2020，64（16）：105–113.

[ 29 ] 辛娟娟，曹佳. 基于复杂网络的文献热点挖掘及可视化 [ J ]. 计算机工程与应用，2016，52（12）：261–264.

[ 30 ] 杨倩. 常见文献计量学工具的分析功能比较研究 [ J ]. 情报探索，2021（10）：87–93.

[ 31 ] 钟辉新. 新兴趋势探测研究综述 [ J ]. 现代情报，2017，37（12）：162–167.

[ 32 ] Holeab C, Paunica M, Curaj A. A complex method of semantic bibliometrics for revealing conceptual profiles and trends in scientific literature. The case of future–oriented technology analysis (FTA) science [ J ]. Economic Computation and Economic Cybernetics Studies and Research, 2017, 51(2): 23–37.

[ 33 ] Wang Y, Luo H, Shi Y. Complex network analysis for international talent mobility based on bibliometrics [ J ]. International Journal of Innovation Science, 2019, 13(11): 419–435.

[ 34 ] Ortega J, Aguillo I. Institutional and country collaboration in an online service of scientific profiles: Google Scholar Citations [ J ]. Journal of Informetrics, 2013, 2(7): 394–403.

[ 35 ] Chae C, Yim J H, Lee J, et al. The bibliometric keywords network analysis of human resource management research trends: the case of human resource management journals in South Korea [ J ]. Sustainability, 2020, 12(14): 5700.

[ 36 ] 何波，赵海媛，袁悦. 中国经理人领域28年研究趋势演变的纵向研究 [ J ]. 重庆大学学报（社会科学版），2016，22（3）：100–108.

[ 37 ] 刘永胜，甘莹莹. 我国绿色食品领域研究现状与趋势分析——基于社会网络的视角 [ J ]. 中国农业资源与区划，2020，41（2）：26–34.

[ 38 ] 杜先芸，高琳霞，徐志营. 基于社会网络分析和共词分析的国内绿色消费行为研究 [ J ]. 电脑编程技巧与维护，2018（6）：20–23.

[ 39 ] 莫富传，娄策群. 高被引论文应用于研究热点识别的理论依据与路径探索 [ J ]. 情报理论与实践，2019，42（4）：59–63.

[ 40 ] 陈贵梧. 图书情报学的国际研究态势：基于2000—2009年SSCI研究性论文的实证分析 [ J ]. 中国图书馆学报，2011，37（1）：73–82.

[ 41 ] 杨爱青，马秀峰，张风燕，等. g指数在共词分析主题词选取中的应用研究 [ J ]. 情报杂志，2012，31（2）：52–55.

[ 42 ] Salton G., McGill M. J. The SMART retrieval system—Experiments in automatic document processing [ J ]. Information Storage and Retrieval, 1971, 6(4): 209–242.

[ 43 ] 宋靖华，龚贤典. 基于复杂网络分析的社区空间网络评价与优化 [ C ]//高等学校建筑学专业教学指导分委员会建筑数字技术教学工作委员会. 数智营造：2020年全国建筑院系建筑数字技术教学与研究学术研讨会论

文集. 长沙：中国建筑工业出版社，2020：39–45.

［44］Newman M, Girvan M. Finding and evaluating community structure in networks［J］. Phys. Rev. E, 2004, 69(2): 026113.

［45］Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks［J］. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): P10008.

# Mining of Research Topics of Green Consumption in China Based on Text Mining and Complex Network

Liu Jieping　Xu Jinya

(Chengdu Neusoft University, Chengdu 611844, China)

---

**Abstract:**［**Purpose/significance**］A fundamental goal of China's 14th Five-Year Plan and Vision 2035 is to promote "green consumption". Identifying research topics is essential because it facilitates staying up-to-date with the latest developments and trends in the field of green consumption, providing indispensable guidance for future studies. ［**Method/process**］Our proposed method comprehensively considers the literature title, abstract, and keywords using text mining and complex network theories. Our method involves utilizing text word segmentation technology to extract subject headings and employing the Analytic Hierarchy Process (AHP) to determine the co-occurrence weight of two-tuple subject phrases. We provide the TI-g index as a proposal by optimizing the traditional word frequency g index through the inclusion of the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm because the word frequency g index is ineffective at filtering out "high-frequency generic words." This study focuses on academic literature on green consumption in the period of 2010 to 2022 in China. ［**Result/conclusion**］A heatmap was generated to display shifts in research topics since 2010, complemented by the identification of recent hotspots of research in this field since 2018. Our analysis identified four major subject fields and highlighted the research limitations present in each.

**Keywords:** Green consumption; Research hotspot; Text mining; Complex network; TF-IDF; Word frequency g index

---

（本文责编：王秀玲）